



## King's Research Portal

DOI:

[10.3366/ijhac.2013.0086](https://doi.org/10.3366/ijhac.2013.0086)

*Document Version*

Early version, also known as pre-print

[Link to publication record in King's Research Portal](#)

*Citation for published version (APA):*

Dunn, S., & Hedges, M. (2013). Crowd-sourcing as a Component of Humanities Research Infrastructures. *International Journal of Humanities and Arts Computing*, 7(1), 147-169. [N/A].  
<https://doi.org/10.3366/ijhac.2013.0086>

### Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

### General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

### Take down policy

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



**Open Access document  
downloaded from King's Research Portal  
<https://kclpure.kcl.ac.uk/portal>**

**Citation to published version:**

Hedges, M., & Dunn, S. (2013). Crowd-sourcing as a Component of Humanities Research Infrastructures. *International Journal of Humanities and Arts Computing*

**This version:** Pre-print

<https://kclpure.kcl.ac.uk/portal/en/publications/crowdsourcing-as-a-component-of-humanities-research-infrastructures%28294ddb42-e10c-4bce-b6cf-8b9500c9aeb7%29.html>

This Pre-print version has been submitted for publication

**The copyright in the published version resides with the publisher.**

**When referring to this paper, please check the page numbers in the published version and cite these.**

**General rights**

Copyright and moral rights for the publications made accessible in King's Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications in King's Research Portal that users recognise and abide by the legal requirements associated with these rights.'

- Users may download and print one copy of any publication from King's Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the King's Research Portal

**Take down policy**

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

# **Crowd-sourcing as a Component of Humanities Research Infrastructures**

**Stuart Dunn, Mark Hedges**

Centre for e-Research, Department of Digital Humanities, King's College London,

26-29 Drury Lane, London, UK

[mark.hedges@kcl.ac.uk](mailto:mark.hedges@kcl.ac.uk), [stuart.dunn@kcl.ac.uk](mailto:stuart.dunn@kcl.ac.uk)

**Abstract:** Crowd-sourcing, the process of leveraging public participation in or contribution to a project or activity, is relatively new to academic research, but is becoming increasingly important as the Web transforms collaboration and communication and blurs the boundaries between the academic and non-academic worlds. At the same time, digital research methods are entering the mainstream of humanities research, and there are a number of initiatives addressing the conceptualisation and construction of research infrastructures for the humanities. This paper examines the place of crowd-sourcing activities within such initiatives, presenting a framework for describing and analysing academic humanities crowd-sourcing, and using this framework of 'primitives' as a basis for exploring potential relationships between crowd-sourcing and humanities research infrastructures.

**Keywords:** crowd-sourcing, research infrastructures, citizen science, scholarly primitives, typology.

## **Introduction**

Crowd-sourcing,<sup>1</sup> the process of leveraging public participation in or contribution to a project or activity, is relatively new to academic research, and even more so to the humanities. However, at a time when the Web is transforming the way in which people collaborate and communicate, and is blurring boundaries between the spaces inhabited by the academic and non-academic worlds, it has never been more

important to examine the role that public communities are beginning to play in academic humanities research.

At the same time, digital research methods are starting to enter the mainstream of humanities research, and there are a number of initiatives addressing the conceptualisation and construction of research infrastructures that would support a shift from ad hoc projects and centres to an environment that is more integrated and sustainable. Such an environment will inevitably be distributed, integrating knowledge, services and people in a loosely-coupled, collaborative 'digital social marketplace'.<sup>2</sup>

The question naturally arises as to where crowd-sourcing activities fit within this framework. More specifically, what contributions can public participants, and the communities to which they belong, make to a humanities research infrastructure, and conversely how can these participants and communities, and the academic researchers who make use of the knowledge and effort that they contribute, benefit from such participation? To begin to address these questions is one of the aims of this paper.

The paper is organised as follows: we begin by describing the context in which the work was carried out, and the methodology used. We then review a number of existing terminologies and typologies for crowd-sourcing and related concepts, and follow this with an analysis of the main motivations for engaging with crowd-sourcing, from both the volunteer's and the academic's points of view. Finally, we build upon this by presenting the outline of a framework for describing and analysing academic humanities crowd-sourcing projects, and use this framework of 'primitives' as a basis for exploring the potential relationships between various forms of crowd-sourcing activity and humanities research infrastructures.

## **Background and Methodology**

The research described in this paper was mostly carried out as part of the *Crowd-sourcing Scoping Study* project (Ref. AH/J01155X/1), which ran for nine months from February-November 2012, and was funded by the Arts and Humanities Research Council as part of its Connected Communities programme.

The study's methodology had four main components:

- a literature review covering academic humanities research that has incorporated crowd-sourcing, research into crowd-sourcing as a method, and less formal outputs such as blogs and project websites.
- two workshops facilitating discussion between, respectively, humanities academics who have used crowd-sourcing, and contributors to crowd-sourcing projects;
- an online survey of contributors to crowd-sourcing projects, exploring their backgrounds, histories of participating in such projects, and motivations for doing so;
- interviews with academics and contributors.

The study does not claim to be comprehensive: there are bound to be important projects, publications, individuals and activities that have been omitted, and there is a strong UK and Anglophone focus on the activities studied. In particular, while the survey was widely publicised, it was self-selecting and makes no claim to being statistically representative; it functioned rather as a means of gathering qualitative information about contributors' backgrounds and motivations.

## Crowd-sourcing and related concepts

The term *crowd-sourcing* was coined in a *Wired* article by Jeff Howe,<sup>3</sup> in which he draws a parallel between reducing labour costs by outsourcing to cheaper countries, and utilising 'the productive potential of millions of plugged-in enthusiasts'. In an academic context, the term has developed from an economic focus to an information focus, in which this productive potential is used to achieve research aims. However, the term is problematic and requires further analysis.

It is first necessary to distinguish crowd-sourcing from some related concepts. It is broader and less easy to define than 'citizen science', which is commonly understood to refer to activities whereby members of the public undertake well-defined and (individually) small-scale tasks as part of larger-scale scientific projects.<sup>4</sup> Another related concept is the 'Wisdom of Crowds',<sup>5</sup> which holds that large-scale collective decision-making can be superior to that of individuals, even experts. Although academic crowd-sourcing *can* be about decision, the decisions involved are rarely as neatly packageable as those implied in the world of business, where the 'good' or 'bad' nature of a decision can be evaluated on the basis of

profitability.<sup>6</sup> Such collective decision-making also lacks the elements of collaboration around activities conceived and directed for a common purpose that characterise crowd-sourcing as commonly understood.

Another important distinction is that between crowd-sourcing and 'social engagement'.<sup>7</sup> According to Holley, social engagement involves 'giving the public the ability to communicate with us and each other', and is 'usually undertaken by individuals for themselves and their own purposes', whereas crowd-sourcing 'uses social engagement techniques to help a group of people achieve a shared, usually significant, and large goal by working collaboratively together as a group'. Holley also notes that crowd-sourcing is likely to involve more effort, and implies a level of commitment and participation that goes beyond casual interest, whereas social engagement is an extension of the kinds of online activities – Tweeting, commenting – that millions do on a daily basis anyway. In one way, this aligns crowd-sourcing with 'citizen science'.

Indeed, Wiggins and Crowston develop this theme by highlighting a distinction between *citizen science* and *community science*, and stating as a key ingredient of the former that it is not self-organising and 'does not represent peer production ... because the power structure of these projects is usually hierarchical'.<sup>8</sup> A fundamental aspect of citizen science is thus that the goal is defined by a particular person or group (almost always as part of a professional academic undertaking), and the participants (recruited through an open call) provide some significant effort towards achieving that goal.

However, the different intellectual traditions of the sciences and the humanities embrace, and are embraced by, different kinds of non-academic community. Indeed, as Trevor Owens has noted, most successful crowd-sourcing activities in the humanities and cultural sectors are not really about crowds at all, in the sense of 'large anonymous masses of people', but are about 'participation from interested and engaged members of the public'.<sup>9</sup> While a crowd-sourcing project may have the capacity for involving large numbers of people, in many cases only a few contributors end up being actively engaged, and these contribute a large percentage of the work. While there may be a centralised recruitment process, at this level the body of contributors is self-organising and self-selecting.

A number of attempts have been made to identify the key characteristics, or to formulate a typology, of crowd-sourcing and related activities. Estellés-Arolas and González-Ladrón-de-Guevara identify eight characteristics, distilled from 32 distinct definitions identified in the literature: the crowd; the task at hand; the recompense obtained; the crowdsourcer or initiator of the crowdsourcing activity; what is obtained by crowdsourcing process; the type of process; the call to participate; and the medium.<sup>10</sup> This extremely processual definition is comprehensive in identifying stages that map easily to business processes.

For the humanities, the 'type of process' is both more significant and more problematic, given the great diversity of processes in the creation of humanities research material. A more task-oriented approach is taken by Wiggins and Crowston,<sup>11</sup> who construct a typology for 'citizen science' activities, identifying five areas of application: Action, Conservation, Investigation, Virtual, and Education. The factors that lead to an activity being assigned to a category are multivariate, and the identification of the categories was based on whether there is an occurrence in a category or not, rather than frequency of those occurrences. The coverage is therefore extremely broad; 'Action', for example, covers self-organising citizen groups that use web technologies to achieve a common purpose, often to do with campaigns on local issues. Moreover, the use of the word 'science' (at least in the usual Anglophone sense) confines the activities reviewed (in terms of both the methods and the content) to a particular epistemic bracket, which inevitably excludes some aspects of humanities research.

One widely-quoted set of definitions for citizen science projects was presented by Bonney *et al.*<sup>12</sup> This divided the field into three broad categories: *contributory projects*, in which members of the public, via an open call, contribute along lines that are tightly defined and directed by scientists; *collaborative projects*, which have a central design but to which members of the public contribute data, and may also help to refine project design, analyze data, or disseminate findings; and *co-created projects*, which are designed by scientists and members of the public working together and for which at least some of the public participants are actively involved in the scientific process. This approach shares important characteristics with the 'task type' described below, in that it is rooted in the complexity of the task, and the amount of initiative and independent analysis required to make a contribution.

The Galleries, Libraries, Archives and Museums (hereafter GLAM) sectors have in particular seen efforts to develop crowd-sourcing typologies. One such typology has been proposed by Mia Ridge in a blog

post,<sup>13</sup> and includes the following categories: Tagging, Debunking (i.e. correcting/reviewing content), Recording a personal story, Linking, Stating preferences, Categorizing, and Creative responses. Again, these categories imply a processual approach, concerning the type of task being carried out, and are potentially extensible across different types of online and physical content and collections.

Another typology from the GLAM domain was developed by Oomen and Aroyo.<sup>14</sup> Their categories include Correction and Transcription, defined as inviting users to correct and/or transcribe outputs of digitisation processes (a category that Ridge's 'Debunking' partially, but not entirely, covers); Contextualisation, or adding contextual knowledge to objects, by constructing narratives or creating User Generated Content (UGC) with contextual data; Complementing Collections, which is the active pursuit of additional objects to be included in a collection; Classification, defined as the gathering of descriptive metadata related to objects in a collection (Ridge's 'Tagging' is a subset of this); Co-curation, which is using inspiration/expertise of non-professional curators to create (Web) exhibits (somewhat analogous to the co-created projects of Bonney *et al.*, but more task-oriented); and Crowdfunding, or the collective cooperation of people who pool their money and other resources together to support efforts initiated by others.<sup>15</sup> Ridge explicitly rejects crowdfunding as a component of crowd-sourcing.<sup>16</sup>

These typologies from the GLAM world perhaps represent best the different crowd-sourcing activities examined by the study, although such lists of categories do not reflect fully the complexity of the situations encountered. Instead, we propose a typology that is orientated along four distinct, although inter-dependent, facets, as described in *Crowd-sourcing and research infrastructures* below.

## **Motivations**

### *Motivations of participants*

#### Overview

Most studies have concluded that crowd-sourcing contributors typically do not have a single motivation; our own survey indicated overwhelmingly (79%) that the contributors who responded have both personal and altruistic motivations. However in many cases it is possible to identify a dominant motivating factor,



which is almost always concerned directly with the activity's subject area. In an analysis of 207 forum posts and interview responses for example, the Galaxy Zoo project found that the top motivations were an interest in astronomy (39%), a desire to contribute (13%) and a concern with the vastness of the universe (11%).<sup>17</sup> A study of volunteers for the Florida Fish and Wildlife Conservation Commission's Nesting Beach Survey found that concern for turtle conservation was the overwhelming motivating factor.<sup>18</sup>

Moreover, studies of the motivations of the contributors to academic crowd-sourcing projects have emphasised personal interest in the subject area concerned, and the opportunities provided to exercise that interest and to engage with people who share it, without material benefit. Such interest is usually concerned with the outcome, but it can also be in the process, or some combination of both. For example, in her 2009 assessment of volunteers to the TROVE project, Holley notes that 'a large proportion was family history researchers', who were highly motivated and had 'a sense of responsibility towards other genealogists to help not only themselves but other people where possible'.<sup>19</sup> In general, it may be said that research into crowd-sourcing motivations suggests a clear primary, although not exclusive, focus on the subject or activity area, and that motivations can be personal or altruistic, and extrinsic or intrinsic.

## Rewards

For the most part, crowd-sourcing projects do not reward their contributors directly in material or professional terms, and conversely contributors to crowd-sourcing projects are not subject to discipline (in either sense) or sanction in the way that members of conventionally-configured research projects are. Indeed, it is clear that the motivations of participants in academic crowd-sourcing tend to be intrinsic to the activity. However, we may regard more indirect benefits as constituting a form of reward: the fulfilment of an interest in the subject; personal gains such as skills, experience or knowledge; some form of status; or a feeling of gratification.

In our survey, contributors mentioned a number of skills gained, including general IT competencies, such as editing wikis and using Skype for distributed collaboration, as well as specialised skills such as TEI encoding. Many contributors gained domain knowledge, for example through the opportunity to edit historical documents (ships' histories) resulting from participation in the Old Weather project. This project showed that the domain interests of the participants can differ from those of the project team, which in this

case is solely interested in those parts of the documents being transcribed that relate to climate history,<sup>20</sup> whereas several contributors became interested in the histories of individual ships, and in addressing niches of history that had been hitherto unexplored. Participants can also pick up a basic grounding in research methods of collation, synthesis and analysis in the area of interest to them.

Less concrete benefits also function as rewards. It was frequently noted that some form of 'feedback loop', through which a participant is informed that their contributions were correct and valuable, is a very important motivating factor for engaging with crowd-sourcing projects, and conversely that a lack of feedback can be very frustrating and discouraging to the participant. Feedback also plays a key role in building a sense of community, and making participants feel that they have a stake in the project. For complex tasks, feedback may also be a necessary part of improving volunteers' work practices, as in Transcribe Bentham.<sup>21</sup> This feedback can be immediate and specific to an individual contribution – for example, participants in the British Library's Georeferencer project (BLG),<sup>22</sup> who could see the results of their work immediately – or it can be deferred and cumulative, for example by means of rankings.

Contributors may receive various 'social' rewards, for example through rankings, increased standing in the crowd-sourcing community, or (in the case of Galaxy Zoo) being credited and named in publications. Similarly, contributors may be subjected to social sanctions, such as banning (e.g. removal of pages or blocking of accounts on Wikipedia), which can adversely affect their reputation and enjoyment, and may even in rare cases reflect on their professional standing.

As well as simple feedback interactions between the project and an individual user, the ability to interact with other participants, for example via a project forum, is an extremely important motivation. Such project-based social networks are used both for 'exchanging chit-chat' and for discussing and sharing information on the practical and technical issues raised, and can foster a sense of community among the participants that can extend beyond the immediate activities of the project itself. A good example of this is the Old Weather forum,<sup>23</sup> which contains exchanges among participants that are indicative of a high degree of collaborative, communal working in addressing problems that arise during the process. The importance of forums was also noted by participants in Transcribing Bentham and British Library Georeferencer.

## Gamification

Some approaches have emphasised the importance of tasks being enjoyable, and have focused on the development of games for crowd-sourcing of different kinds. Prestnoprnik and Crowston discuss the role of games, and in particular possible approaches to creating an application for crowd-sourced natural history taxonomy classification using design science.<sup>24</sup> The Bodiam Castle project provides an example of the potential for games in the context of archaeological analysis of buildings, although this had a greater emphasis on visualisation than on competition.<sup>25</sup>

However, Prestnoprnik and Crowston also note that 'gamification' can act as a disincentive to contributors who have expert knowledge or deep interest in the subject.<sup>26</sup> Gamification can also be a barrier for users who simply want to engage with the assets or processes in question, and can trivialise the process of acquiring or processing data.<sup>27</sup> In their analysis of *The Bird Network* project, in which gathered data about the use of bird-boxes by birds, which was then shared with the scientific team, Brossard *et al.* note that participants' interest in ornithology was likely to overshadow awareness of scientific process,<sup>28</sup> and thus stymie efforts by the Lab to contribute to scientific awareness and education.<sup>29</sup>

## Competition

Although very few participants in our survey admitted to being motivated by competition with each other, among those who attended our workshop competition featured strongly as a factor, although this should be qualified by the fact that those present tended to be 'super contributors', who are likely to feel more competitive than those in the 'long tail' of the crowd.

For many projects it is possible to track individual participant's contributions and to acquire statistics on contributions, and in such cases projects can establish 'leader boards' indicating which participants have made the biggest contributions (in whatever terms the project is using). For example, the British Library's Georeferencer project displayed the handles of the users who processed the most maps, and the 'winner' was invited to meet the Library's head of cartography. The Old Weather project also encouraged competition by assigning roles to contributors based on the number of pages transcribed.

However, in order for competition to be a significant motivating factor, the tasks and their outcomes must be sufficiently quantifiable to allow mutual comparison; matters can become complex when tasks are not comparable directly. For example, in BLG some maps were more complex than others, and the team felt that this affected the meaningfulness of comparing the effort needed to georeference them. Where more creative or interpretive outputs are being created, this lack of commensurability is a still greater issue, and there may even be conflicts between outputs; simple rankings seem inappropriate to such scenarios.

In any case, the encouragement of competition should not be at the cost of alienating potential participants who are not by nature competitive, nor of favouring speed and volume at the expense of quality and care. Indeed, competition can be defined not just in this quantitative sense; volunteers may compete to produce more high-quality work, although in the absence of metrics this can amount to competing only against oneself. Note also that competition is not incompatible with a sense of common purpose; for example, Old Weather participants often ‘feel like part of the ship’ on which they are working.

### *Motivations of academics*

At least part of the success of Galaxy Zoo and other Zooniverse projects is that they catered to clear and present academic needs. In the case of Galaxy Zoo itself, the assets – photographs of galaxies – were far too numerous to be examined individually by any research team, and the task – the classification of those galaxies – was not one that could be performed by computer software, although for the most part could be carried out by a person without specialist expertise.<sup>30</sup> Quite simply, this is work that could not have been carried without large-scale public engagement and participation.

Most cases where humanities academics have engaged with crowd-sourcing have been driven by specific research questions or the need for a particular resource. For example, the Transcribe Bentham project was motivated by the fact that 40,000 folios of Bentham’s work were untranscribed, and thus these valuable primary sources were inaccessible to people researching eighteenth or nineteenth century thought.<sup>31</sup> BLG was motivated by the desire to make its map collections more searchable and thus more exploitable. In Old Weather, researchers were motivated by the desire to be able to use information contained within the assets to explore historic weather patterns, although these motivations may not necessarily be shared by the participants.<sup>32</sup> Although the research motivations are various, the key

characteristic leading the project to use crowd-sourcing is that each involves tasks that a computer could not carry out, and that a research team could only do only with prohibitively large resources.

Note however, during the initial six-month testing period of the project, the rate of volunteer transcription compared unfavourably with that of professional researchers,<sup>33</sup> possibly due to the complexity of the material and the difficulty of Bentham's handwriting. There was also an extremely high moderation overhead, with significant staff time needed to validate the outputs and provide feedback to the contributors. Since then, the volunteer transcription rate has improved significantly, so there is potential for avoiding significant costs in the future.<sup>34</sup> However, this example can serve as a warning against assumptions that crowd-sourcing provides free labour.

Other researchers, particularly those in the GLAM sector, see crowd-sourcing as a means of filling gaps in the coverage of their collections,<sup>35</sup> as it can be an effective way of obtaining information about assets (or the assets themselves) to which only certain members of the public have access, for example through personal or family connections. However, in order to be usable for academic purposes, a degree of curation is required, and this may involve expert input.

It is clear that public engagement and community building is frequently an unintentional by-product of crowd-sourcing projects. In some cases it is seen as an explicit motivation, with the aim of encouraging public engagement with scholarly archives and research, and thus increasing the broader impact of academic research activities.<sup>36</sup>

## Crowd-sourcing and research infrastructures

### *A conceptual framework for crowd-sourcing*

One of the outcomes of our study is a typology for crowd-sourcing in the humanities, which brings together the earlier work cited in Section 2 with the experiences and processes uncovered during the study. It does not seek to provide an alternative set of categories specifically for the humanities, in competition with those considered above. Rather, we propose a model for describing and understanding crowd-sourcing projects in the humanities by analysing them in terms of four key facets – **asset type**,

**process type**, **task type**, and **output type** – and of the relationships between them, and in particular by observing how the applicable categories in one facet are dependent on those in other facets.

**Error! Reference source not found.** illustrates the four facets and their interactions.

- A *process* is composed of *tasks* through which an *output* is produced by operating on an *asset*. It is conditioned by the kind of asset involved, and by the questions that are of interest to project stakeholders (both organisers and volunteers) and can be answered, or at least addressed, using information contained in the asset.
- An *Asset* refers to the content that is, in some way, transformed as a result of processing by a crowd-sourcing activity.
- A *task* is an activity that a project participant undertakes in order to create, process or modify an asset (usually a digital asset). Tasks can differ significantly as regards the extent to which they require initiative and/or independent analysis on the part of the participant, and the difficulty with which they can be quantified or documented. The *task types* were identified the aim of categorising this complexity, and are listed below in approximately increasing order.
- The *output* is what is produced as the result of applying a process to an asset. Outputs can be tangible and/or measurable, but we make allowance also for intangible outcomes, such as awareness or knowledge etc.

**Error! Reference source not found.**–**Error! Reference source not found.** list the categories that the study identified under each facet; these are based for the most part on an examination of existing crowd-sourcing practice, so it is to be expected that the lists will be extended and/or challenged by future work. Detailed descriptions of each category may be found in the report by Dunn and Hedges;<sup>37</sup> in the rest of this paper, we examine the framework specifically in relation to humanities research infrastructures.

### *From crowd-sourcing primitives to research infrastructures*

Rather than attempting to map the elements of this crowd-sourcing framework to specific infrastructures or infrastructural components, we note instead that it may be thought of as a framework of ‘primitives’, in a

sense analogous to that of 'scholarly primitives'. Scholarly primitives may be defined as 'basic functions common to scholarly activity across disciplines',<sup>38</sup> and they provide a conceptual framework for classifying scholarly activities. Given the diversity of humanities research, it is not surprising that there are various sets of candidates – in addition to Palmer *et al.* there are, for example, Unsworth,<sup>39</sup> Benardou *et al.*<sup>40</sup> and Anderson *et al.*<sup>41</sup> – and such a structure has in particular been used as a framework for conceptualising and developing infrastructure for supporting humanities research.<sup>42</sup> The *process* facet in particular may be regarded as providing a set of primitives in this sense, and the output type *composite digital collection with multiple meanings* may in particular be regarded as a form of humanities 'research object', in the sense used by Bechhofer *et al.*<sup>43</sup> and Blanke and Hedges.<sup>44</sup>

Of course, the categorisation into primitives described above is quite different to those in the works cited; this is only to be expected, as it represents the activities of quite different stakeholders, namely interested members of the public rather than professional scholars (although of course one person can play different roles in different circumstances). In particular, there is a greater emphasis on creating or enhancing digital assets in some way, rather than using these assets in research (although again these activities can overlap. For the remainder of this paper, we will look in more detail at each of the *process* types in turn, using specific examples examined by the study with a view to seeing how crowd-sourcing can contribute effectively to humanities research infrastructures.

## **COLLABORATIVE TAGGING**

*Collaborative tagging* may be regarded as crowd-sourcing the organisation of information assets by allowing users to attach tags to those assets. Tags can be based on existing controlled vocabularies, but are more usually derived from free text supplied by the users themselves. Such 'folksonomies' are distinguished from deliberately designed knowledge organisation systems by the fact that they are self-organising, evolving and growing as contributors add new terms. It is possible to extract more formal vocabularies from folksonomies.<sup>45</sup>

Collaborative tagging can result in two concrete outcomes: it can make a corpus of information assets searchable using keywords applied by the user pool, and it can highlight assets that have particular significance, as evidenced by the number of repeat tags they are accorded by the pool. Research in this

area has examined the patterns and information that can be extracted from folksonomies. Golder found that patterns generated by collaborative tagging are, on the whole, extremely stable, meaning that minority opinions can be preserved alongside more highly replicated, and therefore mainstream, concentrations of tags.<sup>46</sup> Other research has shown that user-assigned tags in museums may be quite different from vocabulary terms assigned by curators, and that relating tags to controlled vocabularies can be very problematic,<sup>47</sup> although it could be argued that this allows works to be addressed from a different perspective than that of the museum's formal documentation. In any case, such approaches to knowledge organisation are likely to play a significant part in the organisation of humanities data in the future.

An example is the BBC's YourPaintings project,<sup>48</sup> developed in collaboration with the Public Catalogue Foundation, which has amassed a collection of photographs of all paintings in public ownership in the UK. The public is invited to apply tags to these, which both improves discovery and enables the creation of an aggregation of specialised knowledge.

A more complex example is provided by the Prism project.<sup>49</sup> Collaborative tagging typically assumes that the assets being tagged are themselves stable and clearly identifiable as distinct objects. Prism allowed readers to highlight significant areas of a text and apply tags to them, and thus build up a collective interpretation of the text. Unlike many humanities crowd-sourcing activities, such as transcribing texts according to well-defined procedures, which have identifiable completions, interpretation can go on indefinitely, and there are no right or wrong answers.

## **LINKING**

*Linking* covers the identification and documentation of relationships (usually typed) between individual assets. Most commonly, this takes the form of linking via semantic tags, where the tags describe binary relationships, in which case it is analogous to collaborative tagging. In principle, this could also include the identification of n-ary relationships.

## **TRANSCRIBING**

Transcribing is currently one of the most prominent areas of humanities crowd-sourcing, as it can be used to address a fundamental problem with digitisation, namely the difficulty of rendering handwriting into machine-readable form using current technology. Typically, such transcription requires the human eye



and, in many cases, human interpretation. In terms of our typology, the *output* of a transcribing process will typically be *transcribed text*.

Two projects have contributed significantly to this prominence: *Old Weather* (OW) and *Transcribe Bentham* (TB). OW involved the transcription of ships' log-books held by The National Archives, in order to obtain access to the weather observations they contain, information that is of major significance for climate research.<sup>50</sup> TB encouraged volunteers to transcribe and engage with unpublished manuscripts by the philosopher and reformer Jeremy Bentham, by rendering them into text marked up using TEI XML.<sup>51</sup>

The collaborative model needed for successful crowd-sourced transcription depends on the complexity of the source material. Complex material, such as these two cases, requires a high level of support, whether from the project team or a participant's peers. Simpler material is likely to require less support; for example, when transcribing the more structured data found in family records,<sup>52</sup> the information (text or integers) to be transcribed is presented to the user in small segments – e.g. names, dates, addresses – and transcription requires different cognitive processes that are less dependent on interaction with peers and experts.

Note that this category includes marked-up transcriptions, e.g. using TEI XML, as well as simple transcription of characters. There will be a point however at which the addition of semantic mark-up will go beyond mere transcription, and will count as a form of *collaborative tagging* or *linking*, and the output will typically be *enhanced text*.

## **CORRECTING/MODIFYING CONTENT**

While content is increasingly 'born digital', projects for digitising analogue material abound. Many mass-digitisation technologies, such as Optical Character Recognition (OCR) and speech recognition, can be error-prone, and any such enterprise needs to factor in quality control and error correction, which can make use of crowd-sourcing.

The TROVE project, which produced OCR-ed scans of newspapers from the Australian National Archives, is an excellent example of this.<sup>53</sup> The volume of digitised material precluded the corrections being undertaken by the Archive's its own staff, and using uncorrected text would have significantly reduced the benefits of digitisation, as search capability would have been very restricted.

Another potential application in this category is for correcting automated transcriptions of recorded speech, as such transcription is currently highly error-prone, with error rates of 30% or more.<sup>54</sup>

## **RECORDING AND CREATING CONTENT**

Processes in this category frequently deal with ephemera and intangible cultural heritage. The latter covers any cultural manifestation that does not exist in tangible form; typically, crowd-sourcing is used to document such heritage through a set of processes and tasks, resulting in some form of tangible output. The importance of preserving intangible cultural heritage has been recognised by the UN,<sup>55</sup> and the ways in which this can be documented and curated by distributed communities is an important area for future research.

Frequently this takes the form of a cultural institution soliciting memories from the communities it serves, for example the Tenbury Wells Regal Cinema's *Memory Reel* project.<sup>56</sup> Such processes can incorporate a form of editorial control or *post hoc* digital curation, and their outputs can be edited into more formal publications. Another example is the Scottish Words and Place-names (SWAP) project,<sup>57</sup> which gathered words in Scots, determining which words were in current use and where/how they were used, with the ultimate aim of offering selected words for inclusion in the Scottish Language Dictionaries resource.<sup>58</sup> Candidate words were gathered via the project website as well as via social media – Facebook in particular was an important venue for developing conversations around the material – and words that the project felt were suitable were passed to lexicographers for further scrutiny.

By ephemera, we understand cultural objects that are tangible, but are at risk of loss because of their transitory nature, for example home videos or personal photographs.<sup>59</sup> There are a number of project addressing such assets, for example the Europeana 1914-1918 project,<sup>60</sup> which is collecting digitised personal artefacts relating to the First World War.

The ubiquity of the Web, and access to content creation and digitisation technologies, has led to the creation of non-professionally curated online archives. These have a clear role to play in enriching, augmenting and complementing collections held by memory institutions, and in developing curatorial narratives independent from those of library and archive professionals.<sup>61</sup> Processes in this category are also likely to have elements of the 'social engagement' model, in terms of Holley's distinction.<sup>62</sup>

## COMMENTING, CRITICAL RESPONSES AND STATING PREFERENCES

Processes of this type are likely to count as crowd-sourcing only if there is some specific purpose around which people come together. One example of this is the *Shakespeare's Global Communities* project,<sup>63</sup> which captured audience responses to the 2012 World Shakespeare Festival, with the aim of investigating how 'social networking technologies reshape the ways in which diverse global communities connect with one another around a figure such as Shakespeare'.<sup>64</sup> The question provides a focus for the activity, which, although not itself producing an academic output, provides a dataset for addressing research questions on the modern reception of Shakespeare.

Appropriately managed blogs can provide a platform for focused scholarly interactions of this type. For example, a review by Sonia Massai of *King Lear* on the Year of Shakespeare site attracted controversial responses, leading to an exchange about critical methods as well as content.<sup>65</sup> What differentiates such exchanges from amateur blogging is the scholarly focus and context provided by the project, and its proactive directing of content creation. The project thus provides a tangible link between the crowd and the subject.

## CATEGORISING

Categorising involves assigning assets to predefined categories; it differs from collaborative tagging in that the latter is unconstrained.

## CATALOGUING

Cataloguing – or the creation of structured, descriptive metadata – is a more open-ended process than categorising, but is nevertheless constrained to following accepted metadata standards and approaches. It frequently includes categorising as a sub-activity, e.g. by LoC subject headings.

Cataloguing is a time- and resource-consuming process for many GLAM institutions, and crowd-sourcing has been explored as a means of addressing this. For example, the *What's the Score* project at the Bodleian investigated a cost-effective approach to increasing access to music scores from their collections through a combination of rapid digitisation and crowd-sourcing descriptive metadata.<sup>66</sup>

Cataloguing is related to *contextualising*, as ordering, arraying and describing assets will also make explicit some of their context.

## **CONTEXTUALISING**

Contextualising is typically a more broadly-conceived activity than the related process types of cataloguing or linking, and it involves enriching an asset by adding to it or associating with it other relevant information or content.

## **GEOREFERENCING**

Georeferencing is the process of establishing the location of un-referenced geographical information in terms of a modern coordinate system such as latitude and longitude. Georeferencing can be used to enrich geospatial assets – datasets or texts, including maps, gazetteers or travelogues, that refer to locations on the earth’s surface – that do not include such explicit information.

A major example of crowd-sourcing activity in this area is the British Library Georeferencer project, which aimed to ‘geo-enable’ historical maps in its collections by asking participants to assign spatial coordinates to digitised map images, a task that would have been too labour-intensive for Library staff to undertake themselves. Once georeferenced, the digitised maps are searchable geographically due to the inclusion of latitude and longitude coordinates in the metadata.<sup>67</sup>

## **MAPPING**

Mapping (in the sense of this typology) refers to the process of creating a spatial representation of some information asset(s). This could involve the creation of map data from scratch, but could also be applied to the spatial mapping of concepts, as in a ‘mind map’. The precise sense will depend on the *asset type* to which mapping is being applied.

There is an important distinction between maps and related geospatial assets created by expert organisations, such as the Ordnance Survey, and those created by community-based initiatives. The former may have the authority of a governmental imprimatur, and the distinction of official endorsement. However, the recent emergence of crowd-sourced geospatial assets – a product of the recent global growth in the ownership of hand-held devices with the ability to record location using GPS<sup>68</sup> – has led to

the emergence of resources such as Open Street Map,<sup>69</sup> which has in turn led to a discussion about the reliability of such resources. In general, it has been found that Open Street Map in particular is extremely reliable,<sup>70</sup> but that the specifications for such resources must be carefully defined.<sup>71</sup> The impact of Open Street Map on the cartographic community generally has been noted.<sup>72</sup> The importance of mapping as a means of conveying spatial significance means that this kind of asset is particularly open to different discourses, and possibly conflicting narratives. The digital realm, with its potential for accommodating multiple, diverse, contributions and interpretations, holds great potential for such material.<sup>73</sup>

## **TRANSLATING**

This covers the translation of content from one language to another. In many cases, a crowd-sourced translation will require a strongly collaborative element if it is to be successful, given the semantic interdependencies that can occur between different parts of a text. However, in cases where a large text can be broken up naturally into smaller pieces, a more independent mode of work may be possible; for example, Suda On-Line,<sup>74</sup> which is translating the entries in a 10<sup>th</sup> Century Byzantine lexicon/encyclopaedia. A more modern, although non-academic, example is the phenomenon of ‘fansubbing’, where enthusiasts provide subtitles for television shows and other audiovisual material.<sup>75</sup>

## **Conclusions**

One of the main conclusions of our study is that research involving humanities crowd-sourcing can best be framed and understood through an analysis in terms of four fundamental facets – asset type, process type, task type, and output type – and of the relationships between them. Depending on the activity in question, and what it aims to do, some categories, or indeed some facets, will have primacy. Outputs might be original knowledge, or they might be more ephemeral and difficult to identify: however, considering the processes of both knowledge and resource creation as comprising of these four facets gives a meaningful context to every piece of research, publication and activity we have uncovered in the course of this review. We hope the lessons and good practice we have identified here will, along with this typology, contribute to the development of new kinds of humanities crowd-sourcing in the future.

Significantly, we have determined that most humanities scholars that have used crowd-sourcing as part of some research activity agree that it is not simply a form of ‘cheap labour’ for mass digitisation or resource enhancement; indeed, in a narrowly cost-benefit sense it does not *always* compare well with more conventional mechanisms of digitisation. In this sense, it has truly left its economic roots, as defined by Howe (2006), behind. The creativity, enthusiasm and alternative foci that communities outside that academy can bring to academic research is a resource that is now ripe for tapping into, and the examples above illustrate the rich variety of forms that this tapping can take.

We have noted the similarity between some aspects of our typology and the concept of the ‘scholarly primitive’, which has proved valuable in humanities e-research for providing a conceptual framework of fundamental building blocks for describing scholarly activities and modelling putative research infrastructures for the humanities. We have used this relationship to investigate how crowd-sourcing activities falling under various *process* types can contribute effectively to such research infrastructures.

## Acknowledgements and additional information

A list of the projects investigated by the study, and a description of the survey (including the questions and a summary of the results) may be found in Appendices B and A respectively of (Dunn and Hedges 2012). The project website is at <http://humanitiescrowds.org>, and additional information (in ‘raw’ form) from the workshops organised as part of the study may be found at [http://humanitiescrowds.org/wp-uploads/2012/09/workshop\\_report1.pdf](http://humanitiescrowds.org/wp-uploads/2012/09/workshop_report1.pdf).

We are very grateful to all those who have shared their knowledge and experience with us during the study, and in particular those who agreed to be interviewed, or participated in the workshops, or provided feedback on the project report.

---

<sup>1</sup> We follow the convention of hyphenating ‘crowd-sourcing’; other authors use ‘crowdsourcing’ or ‘crowd sourcing’. In quotations, we preserve the original form.

- 
- <sup>2</sup> T. Blanke, M. Bryant, M. Hedges, A. Aschenbrenner and M. Priddy, 'Preparing DARIAH', *7th IEEE International Conference on e-Science*, Stockholm, Sweden (2011), 158-165,  
<http://dx.doi.org/10.1109/eScience.2011.30>.
- <sup>3</sup> J. Howe, 'The rise of crowdsourcing', *Wired*, 14.06 (2006),  
<http://www.wired.com/wired/archive/14.06/crowds.html>.
- <sup>4</sup> J. Silvertown, 'A new dawn for citizen science', *Trends in ecology & evolution*, 24, No. 9 (2009), 467-71.  
D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky and D. Werthimer, 'SETI@home: an experiment in public-resource computing', *Communications of the ACM*, 45, Issue 11 (2002), 56-61.
- <sup>5</sup> J. Surowiecki, *The wisdom of crowds: why the many are smarter than the few*, 2004.
- <sup>6</sup> D. Brabham, 'Crowdsourcing as a model for problem solving: an introduction and cases', *Convergence: The International Journal of Research into New Media Technologies*, 14, Issue 1 (2008), 75-90.
- <sup>7</sup> R. Holley, 'Crowdsourcing: how and why should libraries do it?', *D-Lib Magazine*, 16, No. 3/4 (2010),  
<http://www.dlib.org/dlib/march10/holley/03holley.html>.
- <sup>8</sup> A. Wiggins and K. Crowston, 'From conservation to crowdsourcing: a typology of citizen science', *System Sciences (HICSS), 2011 44<sup>th</sup> Hawaii International Conference*,  
<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5718708>.
- <sup>9</sup> <http://www.trevorowens.org/2012/05/the-crowd-andthe-library>
- <sup>10</sup> E. Estellés-Arolas and F. González-Ladrón-de-Guevara, 'Towards an integrated crowdsourcing definition', *Journal of Information Science*, 38, No. 2 (2012), 189-200.
- <sup>11</sup> A. Wiggins and K. Crowston, 'From conservation to crowdsourcing: a typology of citizen science'.
- <sup>12</sup> R. Bonney, H. Ballard, R. Jordan, E. McCallie, T. Phillips, J. Shirk and C. Wilderman, *Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science*

---

Education, Center for Advancement of Informal Science Education, Washington D. C. (2009),  
<http://caise.insci.org/uploads/docs/PPSR%20report%20FINAL.pdf>.

<sup>13</sup> <http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.htm>

<sup>14</sup> J. Oomen and L. Aroyo, 'Crowdsourcing in the cultural heritage domain: opportunities and challenges', *Proceedings of the 5<sup>th</sup> International Conference on Communities and Technologies* (2011), 138-149,  
<http://www.cs.vu.nl/~marieke/OomenAroyoCT2011.pdf>.

<sup>15</sup> A. Agrawal, C. Catalini and A. Goldfarb, 'The geography of crowdfunding', *NET Institute Working Paper Series*, 10-8 (2011), 1-57, <http://ssrn.com/abstract=1692661>.

<sup>16</sup> <http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.htm>

<sup>17</sup> M. J. Raddick, G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay and J. Vandenberg, 'Galaxy Zoo: exploring the motivations of citizen science volunteers', *Astronomy Education Review*, 9 (2010), [http://aer.aas.org/resource/1/aerscz/v9/i1/p010103\\_s1](http://aer.aas.org/resource/1/aerscz/v9/i1/p010103_s1).

<sup>18</sup> B. M. Bradford and G. D. Israel, 'Evaluating volunteer motivation for sea turtle conservation in Florida', *Agricultural Education* (2004), 1-9.

<sup>19</sup> R. Holley, *Many hands make light work: public collaborative OCR text correction in Australian historic newspapers*, National Library of Australia (2009),  
[http://www.nla.gov.au/ndp/project\\_details/documents/ANDP\\_ManyHands.pdf](http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf).

<sup>20</sup> <http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf>

<sup>21</sup> T. Causer, J. Tonra and V. Wallace, 'Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham', *Literary and Linguistic Computing*, 27, Issue 2 (2012), 1-19. Similar conclusions were drawn by the authors of the current article, based on their interviews with staff and volunteers from the Old Weather project and the British Library's Georeferencer project.

<sup>22</sup> <http://www.bl.uk/maps/>



---

<sup>23</sup> <http://forum.oldweather.org>

<sup>24</sup> N. R. Prestopnik and K. Crowston, 'Gaming for (citizen) science: exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system', *Proceedings of "Computing for Citizen Science" workshop at the 7<sup>th</sup> IEEE eScience Conference* (2011), [http://crowston.syr.edu/sites/crowston.syr.edu/files/gamingforcitizenscience\\_ver6.pdf](http://crowston.syr.edu/sites/crowston.syr.edu/files/gamingforcitizenscience_ver6.pdf).

<sup>25</sup> <http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Masinton.pdf>

<sup>26</sup> N. R. Prestopnik and K. Crowston, 'Gaming for (citizen) science: exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system' (2011).

<sup>27</sup> See <http://blog.tommorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a> for a combative assertion of this position.

<sup>28</sup> D. Brossard, B. Lewenstein and R. Bonney, 'Scientific knowledge and attitude change: the impact of a citizen science project', *International Journal of Science Education*, 27, Issue 9 (2005), 1029-1121.

<sup>29</sup> D. J. Trumbull, R. Bonney, D. Bascom and A. Cabral, 'Thinking scientifically during participation in a citizen-science project' *Science Education*, 84, Issue 2 (1999), 265-275.

<sup>30</sup> C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. Nichol, A. Szalay, D. Andreescu, P. Murray and J. Vandenberg, 'Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey', *Monthly Notices of the Royal Astronomical Society*, 389, Issue 3 (2008), 1179-1189.

<sup>31</sup> <http://humanitiescrowds.org/wp-uploads/2012/04/Causer.pdf>

<sup>32</sup> <http://humanitiescrowds.org/wp-uploads/2012/04/Brohan.pdf>

<sup>33</sup> T. Causer, J. Tonra and V. Wallace, 'Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham' (2012).

- 
- <sup>34</sup> T. Causer and V. Wallace, 'Building a volunteer community: results and findings from Transcribe Bentham' *Digital Humanities Quarterly*, 6. No. 2 (2012), <http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html>.
- <sup>35</sup> M. Terras, 'Digital curiosities: resource creation via amateur digitisation', *Literary and Linguistic Computing*, 25, No. 4 (2010), 425-438, doi:10.1093/lc/fqq019.
- <sup>36</sup> M. Moyle, J. Tonra and V. Wallace, 'Manuscript transcription by crowdsourcing: Transcribe Bentham'. *Liber Quarterly - The Journal of European Research Libraries*. 20, Issue 3/4 (2011).
- <sup>37</sup> S. Dunn and M. Hedges, 'Crowd-sourcing scoping study: engaging the crowd with humanities research', Arts and Humanities Research Council report (2012), <http://humanitiescrowds.org/wp-uploads/2012/12/Crowdsourcing-connected-communities.pdf>.
- <sup>38</sup> C. L. Palmer, L. C. Teffeuau and C. M. Pirmann, 'Scholarly information practices in the online environment: themes from the literature and implications for library service development' (2009).
- <sup>39</sup> J. Unsworth, 'Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this', *'Humanities Computing, Formal Methods, Experimental Practice' Symposium, King's College London* (2000), <http://people.lis.illinois.edu/~unsworth/Kings.5-00/primitives.html>.
- <sup>40</sup> A. Benardou, P. Constantopoulos, C. Dallas and D. Gavrilis, 'Understanding the information requirements of arts and humanities scholarship', *International Journal of Digital Curation*, 5, No. 1 (2010), 18-33.
- <sup>41</sup> S. Anderson, T. Blanke and S. Dunn, 'Methodological commons: arts and humanities e-science fundamentals', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368, No. 1925 (2010), 3779-3796.
- <sup>42</sup> T. Blanke and M. Hedges, 'Scholarly primitives: building institutional infrastructure for humanities e-science', *Future Generation Computer Systems*, 29, Issue 2 (2013), 654-661, <http://dx.doi.org/10.1016/j.bbr.2011.03.031>.

- 
- <sup>43</sup> S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi and C. Goble, *Future Generation Computer Systems*, 29, Issue 2 (2013), 599–611, <http://dx.doi.org/10.1016/j.future.2011.08.004>.
- <sup>44</sup> T. Blanke and M. Hedges, 'Scholarly primitives: building institutional infrastructure for humanities e-science' (2013).
- <sup>45</sup> H. Lin and J. Davis, 'Computational and crowdsourcing methods for extracting ontological structure from folksonomy', *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, 6089 (2010), 472-477, DOI:10.1007/978-3-642-13489-0\_46.
- <sup>46</sup> S. Golder, 'Usage patterns of collaborative tagging systems', *Journal of Information Science*, 32, Issue 2 (2006), 198-208.
- <sup>47</sup> J. Trant, *Tagging, Folksonomy, and Art Museums: Results of steve.museum's research* (2009), [http://conference.archimuse.com/blog/jtrant/stevemuseum\\_research\\_report\\_available\\_tagging\\_for](http://conference.archimuse.com/blog/jtrant/stevemuseum_research_report_available_tagging_for); J. Trant, D. Bearman and S. Chun, 'The eye of the beholder: steve.museum and social tagging of museum collections', *Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM07)*, Toronto, Canada (2007).
- <sup>48</sup> <http://www.bbc.co.uk/arts/yourpaintings/>
- <sup>49</sup> <http://www.scholarslab.org/category/praxis-program/>
- <sup>50</sup> P. Brohan, R. Allan, J. E. Freeman, A. M. Waple, D. Wheeler, C. Wilkinson and S. Woodruff, 'Marine observations of old weather' *Bulletin of the American Meteorological Society*, 90, Issue 2 (2009), 219-230.
- <sup>51</sup> T. Causer, J. Tonra and V. Wallace, 'Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham' (2012).
- <sup>52</sup> For example, <http://www.familysearch.org>

---

<sup>53</sup> R. Holley, *Many hands make light work: public collaborative OCR text correction in Australian historic newspapers* (2009).

<sup>54</sup> M. Wald, 'Crowdsourcing correction of speech recognition captioning errors' *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility - W4A '11* (2011), <http://eprints.soton.ac.uk/272430/1/crowdsourcercaptioningw4allCRv2.pdf>.

<sup>55</sup> R. Kurin, 'Safeguarding intangible cultural heritage in the 2003 UNESCO convention: a critical appraisal', *Museum International*, 56, Issue 1-2 (2004), 66–77.

<sup>56</sup> <http://www.regaltenbury.org.uk/memory-reel/>

<sup>57</sup> C. Hough, E. Bramwell and D. Grieve, *Scots Words and Place-Names Final Report*, JISC (2011), <http://www.jisc.ac.uk/media/documents/programmes/digitisation/swapfinalreport.pdf>. See also <http://swap.nesc.gla.ac.uk/>.

<sup>58</sup> <http://www.scotsdictionaries.org.uk/>

<sup>59</sup> This usage differs from the standard usage of the term by museums.

<sup>60</sup> <http://www.europeana1914-1918.eu/en/contributor>

<sup>61</sup> M. Terras, 'Digital curiosities: resource creation via amateur digitisation' (2010).

<sup>62</sup> R. Holley, 'Crowdsourcing: how and why should libraries do it?' (2010).

<sup>63</sup> [www.yearofshakespeare.com](http://www.yearofshakespeare.com)

<sup>64</sup> [http://humanitiescrowds.org/wp-uploads/2012/09/workshop\\_report1.pdf](http://humanitiescrowds.org/wp-uploads/2012/09/workshop_report1.pdf)

<sup>65</sup> <http://bloggingshakespeare.com/year-of-shakespeare-king-lear-at-the-almeida>

<sup>66</sup> <http://www.whats-the-score.org>; <http://scores.bodleian.ox.ac.uk>

<sup>67</sup> C. Fleet, K. C. Kowal and P. Pridal, 'Georeferencer: crowdsourced georeferencing for map library collections', *D-Lib Magazine*, 18, No. 11/12 (2012), <http://www.dlib.org/dlib/november12/fleet/11fleet.html>.

---

<sup>68</sup> M. Goodchild, 'Editorial: citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0', *International Journal of Spatial Data Infrastructures Research*, 2 (2007), 24-32.

<sup>69</sup> <http://www.openstreetmap.org/>

<sup>70</sup> M. Haklay and P. Weber, 'OpenStreetMap: user-generated street maps', *Pervasive Computing, IEEE*, 7, Issue 7 (2008), 12-18. M. Haklay, 'How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets', *Environment and Planning B: Planning and Design*, 37, Issue 4 (2010), 682-703.

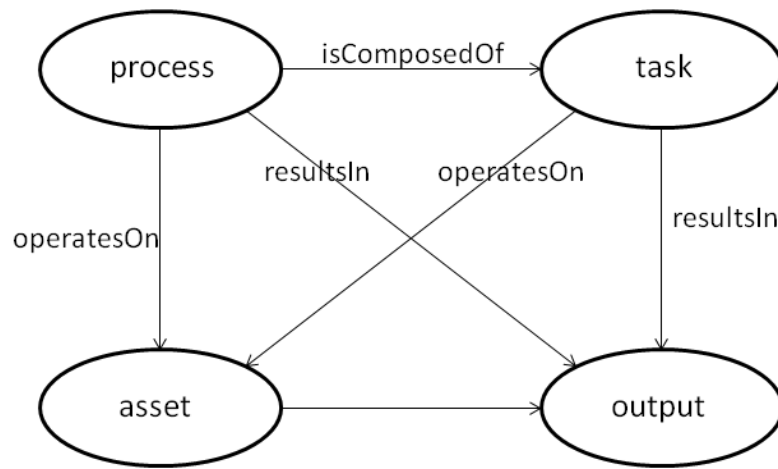
<sup>71</sup> C. Brando and B. Bucher, 'Quality in user generated spatial content: a matter of specifications', *Proceedings of the 13<sup>th</sup> AGILE International Conference on Geographic Information Science*, Guimarães, Portugal (2010), 1-8.

<sup>72</sup> S. Chilton, 'Crowdsourcing is radically changing the geodata landscape: case study of OpenStreetMap', *Proceedings of the 24th International Cartographic Conference*, Santiago, Chile (2009), [http://w.icaci.org/files/documents/ICC\\_proceedings/ICC2009/html/nonref/22\\_6.pdf](http://w.icaci.org/files/documents/ICC_proceedings/ICC2009/html/nonref/22_6.pdf).

<sup>73</sup> C. Fink, 'Mapping together: on collaborative implicit cartographies, their discourses and space construction', *Journal for Theoretical Cartography*, 4 (2011), 1-14. M. Graham, 'Neogeography and the palimpsests of place: Web 2.0 and the construction of a virtual earth', *Tijdschrift voor Economische en Sociale Geografie*, 101, Issue 4 (2010), 422-436.

<sup>74</sup> <http://www.stoa.org/sol/>

<sup>75</sup> J. D. Cintas and P. M. Sanchez, 'Fansubs: audiovisual translation in an amateur environment', *Journal of Specialised Translation*, 6 (2006), 37-52.



**Figure 1: Typology framework**

Process Type
Collaborative tagging
Linking
Correcting/modifying content
Transcribing
Recording and creating content
Commenting, critical responses and stating preferences
Categorising
Cataloguing
Contextualisation
Mapping
Georeferencing
Translating

**Table 1: Process Types**

Asset Type
Geospatial
Text
Numerical or statistical information
Sound
Image
Video
Ephemera and intangible cultural heritage

**Table 2: Asset Types**



TASK
Mechanical
Configurational
Editorial
Synthetic
Investigative
Creative

**Table 3: Task Types**

Asset Type
Original text
Transcribed text
Corrected text
Enhanced text
Transcribed music
Metadata
Structured data
Knowledge/awareness
Funding
Synthesis
Composite digital collection with multiple meanings

**Table 4: Output Types**